Capstone Final Report

KDD Cup 1998

1. **Introduction**
2. **Data Source**

This project uses the dataset in The Second International Knowledge Discovery and Data Mining Tools Competition, held in conjunction with KDD-98.

The dataset was provided by the Paralyzed Veterans of America (PVA), a not-for-profit organization that provides programs and services for US veterans with spinal cord injuries or disease. PVA is also one of the largest direct mail fund raisers in the country. The analysis dataset includes:

* A subset of the 3.5 million donors who received fund raising appeals from PVA in 1997.
* A flag to indicate respondents to the appeal and the dollar amount of their donation
* PVA promotion and giving history
* Overlay demographics, including a mix of household and area level data.

This project uses the Training data set in the competition, which includes 95412 rows and 481 columns.

* Column TARGET\_B: a target variable, binary variable indicating whether or not the record responded to the promotion of interest (‘97NK’ mailing)
* Column TARGET\_D: contains donation amount (dollar), and is only observed for those that responded to the promotion.

Note that I made my training and test sets out of this training file. Thus my training and test sets combined have 95412 cases.

1. **Project objectives**

The objectives of this project are two-fold:

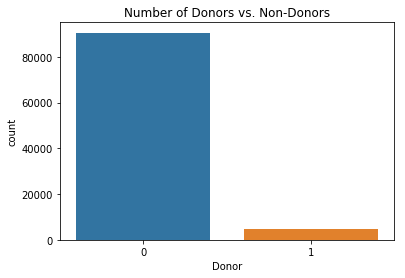
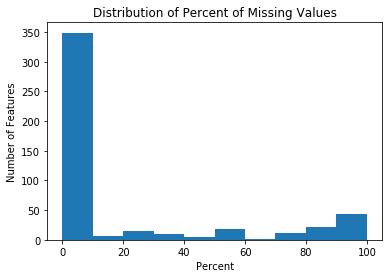
* Predict who will respond to the 97NK mailing promotion (a classification problem)
* For those who are predicted to be donors, predict the amount of their donation (a regression problem).

1. **Data Munging**
2. **Challenges with the data set:**

* The data set is highly imbalanced – it contains only 5% of donors.

This creates an issue for classification models, as the positive records are overpowered by the negative records, thus may cause all predictions to be negative (in the case of Logistic Regression).

* About a quarter of the columns have more than 30% missing values. Imputing the missing values for these columns might introduce bias, thus I decided to drop all these columns.
* The data set has too many features, thus feature selection is crucial in order to have a good model.

1. **Data Munging Tasks**

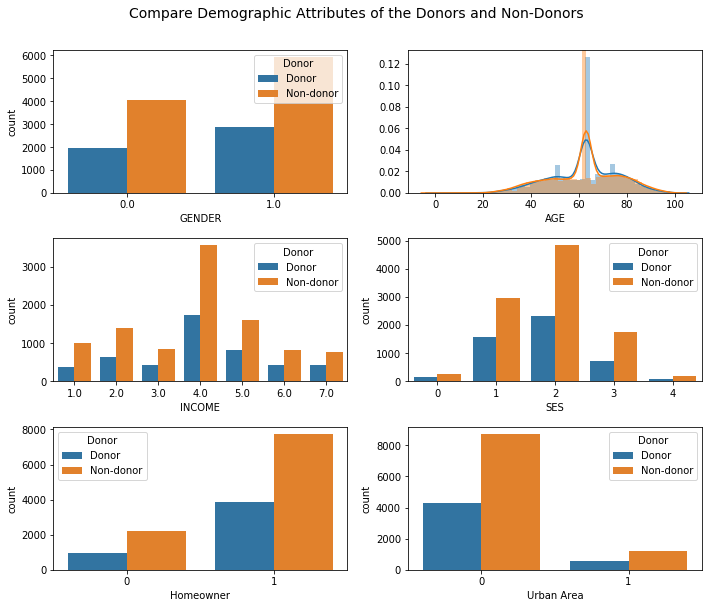
* I created a balanced subset of the raw data set to include all records with positive labels (~4800 records) and a random subset of records with negative labels (~10000 records).

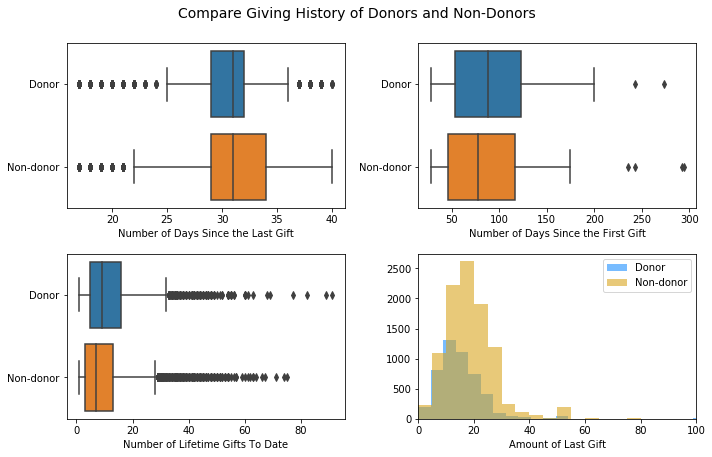
The share of donors in the new data set becomes 32%.

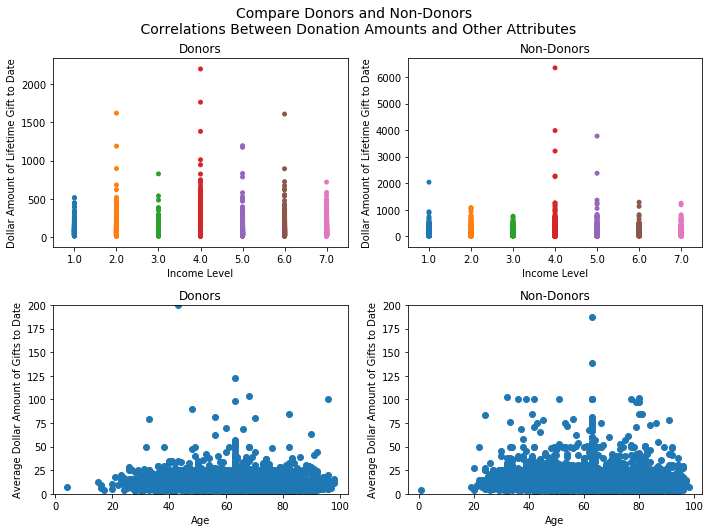
* Checked columns with missing values and dropped those with >30% missing values. This results in 112 columns being dropped.
* Dropped some variables which are believed to be unimportant
* Split some categorical variables which were coded as matrix into components
* Imputed missing values with the mode for numeric variables, and with the most frequent value for categorical variables
* Created some new variables based on existing ones, i.e., some date columns.
* Created dummy variables from categorical variables.
* At the end of this cleaning process, my data set has 14,843 rows and 351 features.

1. **Exploratory Data Analysis**

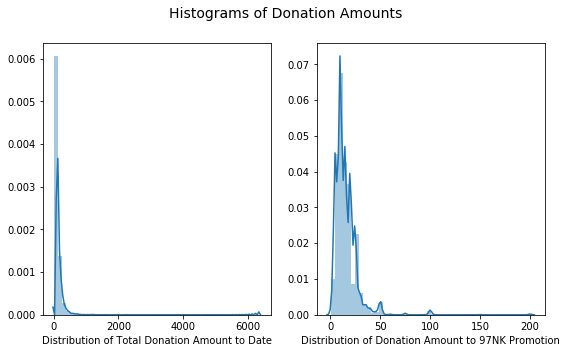
* Using the above cleaned data set, I did some EDA to compare donors and non-donors. It appears that donors and non-donors have very similar demographic attributes (such as age, gender, income, socio-economic status) as well as giving history (last date and first date of donation, total number of donations to date, and amount of most recent donation). It’s not easy to separate donors and non-donors by looking at these attributes.







* The target variable (donation amount to the promotion of interest) range from 0 to 200. Most donors make quite small donations.



1. **Feature Selection For Classification Task**

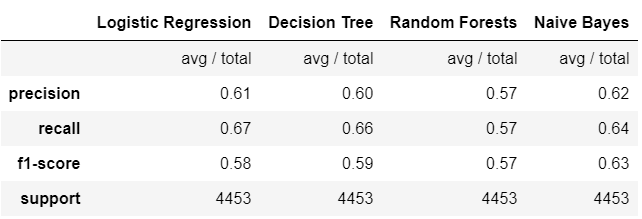
* First, I normalized the data set to prepare for feature selection using variance threshold. Setting a threshold of 0.05 results in 52 features being selected from 351 features.
* Then I performed further feature selection using recursive feature elimination (RFE). This step further reduced the number of features from 52 to 26.
* Final model includes 26 features as below:

['INCOME', 'GENDER', 'DATASRCE', 'POP90C1', 'POP90C2', 'POP90C3', 'HVP1', 'HC2', 'HC7', 'HC8', 'RFA\_2F', 'HOMEOWNR\_H', 'HOMEOWNR\_U', 'RFA\_2A\_D', 'RFA\_2A\_E', 'RFA\_2A\_F', 'RFA\_2A\_G', 'GEOCODE2\_A', 'GEOCODE2\_B', 'GEOCODE2\_C', 'GEOCODE2\_D', 'URBAN\_C', 'URBAN\_R', 'URBAN\_S', 'URBAN\_T', 'URBAN\_U']

* I also tried principal component analysis, but it didn’t produce a better set of features (and better model performance), plus interpretation would be harder. Thus I decided not to go with this option.

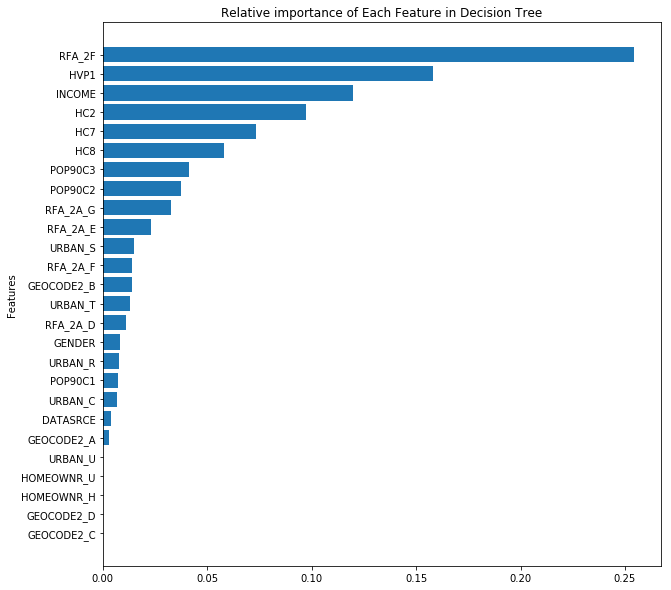
1. **Task I: Predict Who Will be Donors**

* First, I split the data set into training and test sets (70/30).
* First, I split Then I tried several classifiers (Logistic Regression, Decision Trees, Naïve Bayes, Random Forests) on the reduced data set. The results show that overall F1 scores range from 0.57 to 0.63 across all models. Naïve Bayes seems to be the best performing model, followed by Decision Tree.

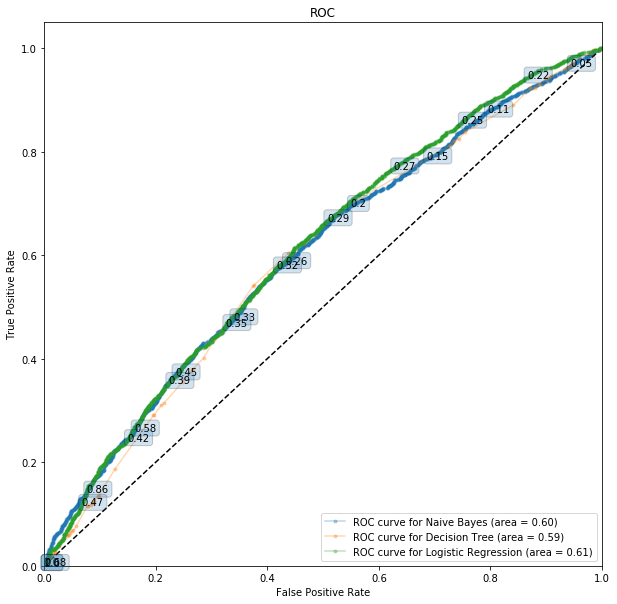


* Features of the highest importance predicted by Decision Tree include: RFA\_2F (Frequency of donations in most recent period), HVP1 (Percent home value >200k), and

Income.



* Naïve Bayes, Decision Tree, and Logistic Regression also have very similar ROC curves and area under the curve (AUC).

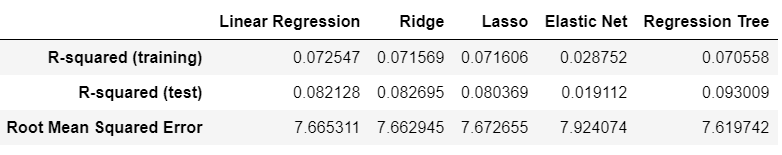


1. **Task II – Predict donation amounts for those who are predicted to be donors**

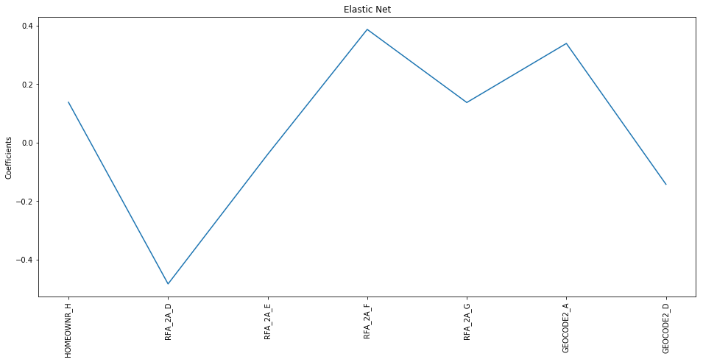
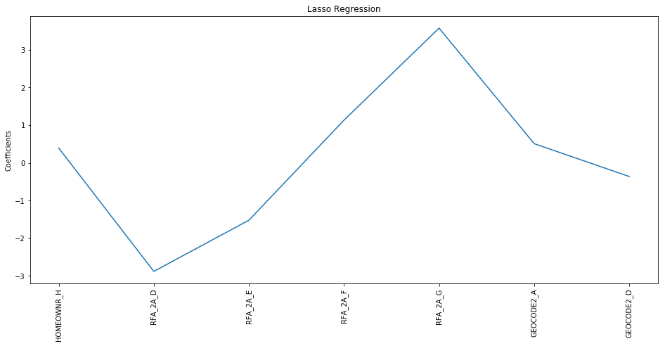
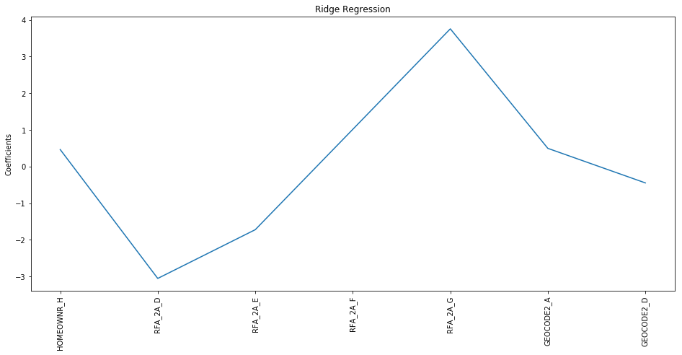
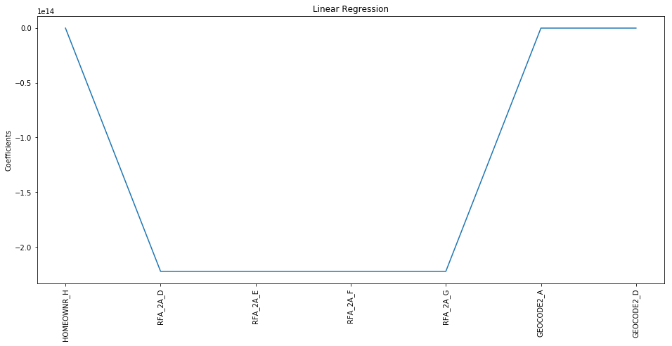
* Using only the samples who are predicted to be donors by the Naïve Bayes model in the first task, I continued to predict the amount of their donations. There are totally 3481 people who are predicted to be donors by Naïve Bayes.
* First, I performed feature selection again using Recursive Feature Elimination, in order to select the best features needed for this regression task (Since the above 26 features were selected specifically for classification task, with Logistic Regression and roc\_auc as arguments for estimator and score in the RFECV function, I decided to look for another set of features that would be useful specifically for the regression task). Thus, I performed RFE on the same 52 features selected by Variance Threshold in task 1, this time with the estimator = linear regression and score func = ‘R2’ as arguments. This step gave me 7 features as below:

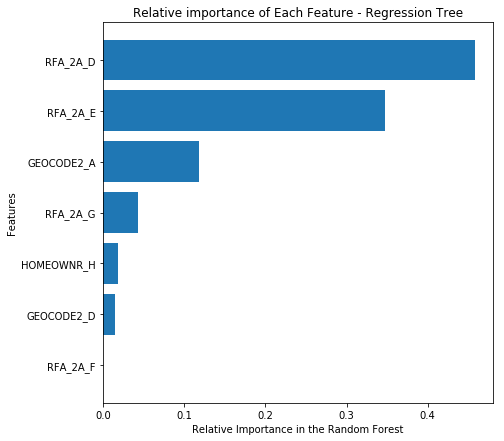
['HOMEOWNR\_H', 'RFA\_2A\_D', 'RFA\_2A\_E', 'RFA\_2A\_F', 'RFA\_2A\_G', 'GEOCODE2\_A', 'GEOCODE2\_D']

* I tried several regression models (Linear Regression, Ridge, Lasso, Elastic Net, Regression Tree), and all of them have very low R-squared, ranging from 0.02 (Elastic Net) to 0.09 (Regression Tree).



* Amount of most recent donation (RFA\_2A\_D & RFA\_2A\_G) are predicted to be the most important features by all models.





1. **Concluding Remarks**

* Feature Selection is probably the most important part of this project. But even if I tried different methods for feature selection, I couldn’t spot an optimal set of variables for both the classification and regression tasks, which is vital for good performance.
* Creating a balanced subset was very useful, since otherwise all of my classifiers would not be able to predict any positive cases due to the overwhelming number of negative cases.
* In the regression task, my R-squared in all models is very low. However, I found that this has to do with the way I created the balanced subset. i.e., if I included an equal number of non-donors and donors (i.e., the ratio of donors vs. non-donors is 1:1), then my regression models would perform much better (R-squared will increase to around 70%).

1. **Acknowledgement**

I would like to give special thanks to my mentor, Jamin Atkins, who gave me so much advice and guidance throughout the whole process. Without his tremendous support and patience, this project would be much more difficult to complete.